Lois de probabilités liées aux tirages de boules dans une urne
Approche sondage : échantillonnage et estimation dans une population finie

Dans le nouveau programme de seconde, rentrée 2009, sont inscrites les notions d'intervalle de fluctuation d'une fréquence d'échantillonnage et d'estimation d'une proportion par intervalle de confiance.

L'échantillonnage et l'estimation dans une population finie (c'est l'objet de la théorie des sondages) fournit une introduction très concrète à la statistique inférentielle. C'est l'objet de ce document.

La population est représentée par une urne dans laquelle les boules sont discernées (paragraphe 1) et dans laquelle il y a deux catégories de boules (paragraphe 2). Dans ce paragraphe, on s'intéresse, parmi les boules tirées, au nombre de boules d'une certaine catégorie, dite catégorie d'intérêt. Selon le nombre et le mode de tirage des boules dans l'urne, on obtient la loi de probabilité de Bernoulli, binomiale ou hypergéométrique. Dans le paragraphe 3, on en déduit la loi de probabilité de la fréquence de boules de la catégorie d'intérêt et on utilise les approximations des lois de probabilités par des lois plus faciles à manier (cf. *tableaux des lois de probabilités usuelles et de leurs approximations*) pour répondre à quelques questions. Enfin, dans le paragraphe 4, on donne quelques éléments de statistique inférentielle (intervalle de fluctuation et intervalle de confiance).

## 1. Questions préliminaires. Probabilités d'événements liés à différents modes de tirages.

**Une urne contient *N* boules numérotées de 1 à *N*.**

**(i) On tire "au hasard" *une* boule de l'urne. Calculer la probabilité de tirer la boule *k*, $k \in [\![1, N]\!]$.**

Espace probabilisé : équiprobabilité sur $[\![1, N]\!]$ ; la probabilité de tirer la boule *k* est 1/*N*

**(ii) On tire "au hasard" *successivement* et *avec remise n* boules de l'urne.**
**Pour tout *k* de $[\![1, N]\!]$ et tout *i* de $[\![1, n]\!]$, calculer la probabilité de l'événement $A_{i,k}$ = "la boule *k* sort au *i*ème tirage", de l'événement $A_k$ "la boule k sort au moins une fois lors des *n* tirages".**
Espace probabilisé : équiprobabilité sur l'ensemble des *n*-listes de $[\![1, N]\!]$ de cardinal $N^n$ ;
$P(A_{i,k}) = N^{n-1} / N^n = 1/N$ et $P(A_k) = 1-(N-1)^n / N^n = 1-(1-1/N)^n$, probabilité approchée par *n* / N dans le cas où *n* est petit devant *N*.

**(iii) On tire "au hasard" *successivement* et *sans remise n* boules de l'urne ($1 \le n \le N$).**
**Pour tout *k* de $[\![1, N]\!]$ et tout *i* de $[\![1, n]\!]$, calculer la probabilité de l'événement $A_{i,k}$ = "la boule *k* sort au *i*ème tirage", de l'événement $A_k$ "la boule k sort lors des n tirages".**

Espace probabilisé : équiprobabilité sur l'ensemble des arrangements de *n* éléments de $[\![1, N]\!]$ de cardinal $A_N^n$ ; $P\left(A_{i,k}\right) = A_{N-1}^{n-1} / A_N^n = 1/N$ et $P\left(A_k\right) = \sum_{i=1}^{n} P\left(A_{i,k}\right) = n/N$ car les événements $\left(A_{i,k}\right)_{i=1,\ldots,n}$ sont deux à deux incompatibles.

**(iv) On tire "au hasard" *simultanément* *n* boules de l'urne ($1 \le n \le N$). Calculer la probabilité de tirer la boule *k*, $k \in [\![1, N]\!]$.**

Espace probabilisé : équiprobabilité sur l'ensemble des combinaisons de *n* éléments de $[\![1, N]\!]$ de cardinal $\binom{N}{n}$ ; $P\left(A_k\right) = \binom{N-1}{n-1} / \binom{N}{n} = n/N$ **.**

## 2. Loi de probabilité du nombre de boules de la catégorie d'intérêt selon divers modes de tirages.

**Une urne contient *N* boules dont *K* rouges (0<*K*<*N*) (catégorie d'intérêt).**

Dans la suite, afin de les dénombrer, on distingue encore les boules de l'urne.
L'urne est alors représentée par l'ensemble $[\![1, N]\!]$ ; on note *R* le sous-ensemble des boules rouges de l'urne et *p* la *proportion* (appelée aussi *fréquence* dans des contextes similaires) de boules rouges dans l'urne avant tirage : $p = K / N$ .

**(i) On tire "au hasard" *une* boule de l'urne. Calculer la probabilité de tirer une boule rouge.**

Espace probabilisé : équiprobabilité sur $[\![1, N]\!]$ ; la *probabilité* de tirer une boule rouge est alors la *proportion p* de boules rouges dans l'urne. Il s'agit du *premier lien entre fréquence et probabilité* ; on l'appellera *approche classique* ou *laplacienne* de la probabilité. C'est Laplace qui, le premier, a écrit que la probabilité d'un événement est le nombre d'issues favorables sur le nombre d'issues possibles à condition que toutes ces issues aient les mêmes chances d'apparaître.

**(ii) On tire "au hasard" *successivement* et *avec remise n* boules de l'urne.**
**Soit $X_i$ ($i \in [\![1, n]\!]$) la v.a. indicatrice de l'événement "la *i*ème boule tirée est rouge" et *X* le nombre de boules rouges tirées. Donner la loi de probabilité de $X_i$ ($i \in [\![1, n]\!]$), de *X* .**

Les variables aléatoires réelles $X_i$ ($i \in [\![1, n]\!]$) et $X$ sont définies sur l'espace probabilisé de la question I.(ii) : équiprobabilité sur l'ensemble des *n*-listes de $[\![1, N]\!]$ de cardinal $N^n$ ;
en reprenant les notations de la question I, on a, pour $i \in [\![1, n]\!]$,

$P\left(\left[X_i = 1\right]\right) = P\left(\bigcup_{k \in R} A_{i,k}\right) = \sum_{k \in R} P\left(A_{i,k}\right) = K/N = p$ ; $P\left(\left[X_i = 0\right]\right) = 1 - p$ ;

$X_i$ ($i \in [\![1, n]\!]$) suit une loi de Bernoulli de paramètre *p*.

Pour $j \in [\![0, n]\!]$, $P\left(\left[X = j\right]\right) = \binom{n}{j} \dfrac{K^j \left(N-K\right)^{n-j}}{N^n} = \binom{n}{j} p^j \left(1-p\right)^{n-j}$ car l'événement $\left[X = j\right]$ est l'ensemble de toutes les *n*-listes de $[\![1, N]\!]$ contenant *j* éléments de *R* et *n* - *j* éléments du complémentaire de *R* (éléments non nécessairement distincts) ; si l'on ne tient pas compte des

numéros de boules, le nombre de ces listes est $\binom{n}{j}$ égal au nombre de position des *j* rouges dans la *n*-liste ; en distinguant les boules il y a $K^j (N-K)^{n-j}$ façons de choisir chacune d'elles d'où le résultat ; la variable aléatoire *X* suit une loi binomiale de paramètres *n* et *p*.
On notera que les probabilités sont calculées en utilisant la formule "nombre de cas favorables sur nombre de cas possibles" puisque nous avons proposé des modèles probabilisés avec équiprobabilité. Nous n'utilisons pas de probabilités conditionnelles ou d'indépendance en probabilités. On utilise l'indice *j* pour ne pas le confondre avec l'indice *i* (pour les tirages) ou avec l'indice *k* (pour les éléments de l'urne) déjà utilisés.

**(iii) On tire "au hasard" *successivement* et *sans remise n* boules de l'urne ($1 \le n \le N$ ).**
**Soit $X_i$ ($i \in [\![1,n]\!]$) la v.a. indicatrice de l'événement "la *i*ème boule tirée est rouge" et *X* le nombre de boules rouges tirées. Donner la loi de probabilité de $X_i$ ($i \in [\![1,n]\!]$), de *X*.**

Les variables aléatoires réelles $X_i$ ($i \in [\![1,n]\!]$) et $X$ sont définies sur l'espace probabilisé de la question I.(iii) : équiprobabilité sur l'ensemble des arrangements de *n* éléments de $[\![1,N]\!]$ de cardinal $A_N^n$ ; en reprenant les notations de la question I, on a, pour $i \in [\![1,n]\!]$,

$P([X_i = 1]) = P\left(\underset{k\in R}{\cup} A_{i,k}\right) = \sum_{k\in R} P(A_{i,k}) = K/N = p$ ; $P([X_i = 0]) = 1-p$ ;

$X_i$ ($i \in [\![1,n]\!]$) suit une loi de Bernoulli de paramètre p.

Soit $j \in [\![0,n]\!]$ ; si $j > K$ ou $n - j > N - K$, alors $P([X = j]) = 0$. Soit $j \in [\![0,n]\!]$ tel que $j \le K$ et $n - j \le N - K$ alors :

$$P([X = j]) = \binom{n}{j} \frac{K(K-1)...(K-j+1)(N-K)(N-K-1)...(N-K-n+j+1)}{N(N-1)...(N-n+1)} = \frac{\binom{K}{j}\binom{N-K}{n-j}}{\binom{N}{n}}$$

car l'événement $[X = j]$ est l'ensemble de tous les arrangements de *n* éléments de $[\![1,N]\!]$ contenant *j* éléments distincts de *R* et *n* - *j* éléments distincts du complémentaire de *R* ;
si l'on ne tient pas compte des numéros de boules, le nombre de ces arrangements est $\binom{n}{j}$ égal au nombre de position des *j* rouges dans le *n*-uplet ; en distinguant les boules il y a $K(K-1)...(K-j+1)(N-K)(N-K-1)...(N-K-n+j+1)$ façons de choisir chacun d'eux d'où le résultat ; la variable aléatoire *X* suit une loi hypergéométrique de paramètres, *N*, *n* et *p* (avec $p = K/N$, proportion de boules rouges dans l'urne avant tirage).

**(iv) On tire "au hasard" *simultanément n* boules de l'urne ($1 \le n \le N$ ).**
**Soit *X* le nombre de boules rouges tirées. Donner la loi de probabilité de *X*.**
La variable aléatoire réelle X est définie sur l'espace probabilisé de la question I.(iv) : équiprobabilité sur l'ensemble des combinaisons de *n* éléments de $[\![1,N]\!]$ de cardinal $\binom{N}{n}$.

Pour $j \in [\![0,n]\!]$, $P([X = j]) = \frac{\binom{K}{j}\binom{N-K}{n-j}}{\binom{N}{n}}$ car l'événement $[X = j]$ est l'ensemble de tous les sous-ensembles de *n* éléments de $[\![1,N]\!]$ contenant *j* éléments de *R* et *n* - *j* éléments du complémentaire de *R* ; il y a $\binom{K}{j}\binom{N-K}{n-j}$ façons de choisir un tel sous-ensemble d'où le résultat.
La variable aléatoire *X* suit une loi hypergéométrique de paramètres *N*, *n* et *p*.

## 3. Loi de probabilité de la fréquence d'échantillonnage et approximation

**On pose $F = X / n$ (fréquence ou proportion de boules rouges tirées dans l'échantillon). Quelle est la loi de probabilité de *F* dans chacun des tirages (ii) (iii) (iv) ? Donner son espérance mathématique et son écart-type.**
**Écrire une loi de probabilité approchée de *F* utilisable dans les applications.**
**Application numérique**
**a) Dans le cas *N* = 10000, *K* = 3000, *n* = 20, calculer** $P(0.2 \le F \le 0.4)$.
**b) Dans le cas *N* = 10000, *K* = 3000, *n* = 200, calculer** $P(0.2 \le F \le 0.4)$.
**c) Dans le cas *N* = 10000, *K* = 500, *n* = 20, calculer** $P(X > 1)$.

(ii) La loi de probabilité de *F* se déduit de celle de *X* qui est binomiale de paramètres *n*, *p*. Pour $j \in [\![0, n]\!]$, $P([F = j/n]) = P([X = j]) = \binom{n}{j} p^j (1-p)^{n-j}$ ; comme on a $E(X) = np$ et $\sigma(X) = \sqrt{np(1-p)}$, on en déduit $E(F) = p$ et $\sigma(F) = \sqrt{p(1-p)/n}$.

(iii) et (iv) La loi de probabilité de *F* se déduit de celle de *X* qui est hypergéométrique de paramètres *N*, *n*, *p*. Pour $j \in [\![0, n]\!]$, $P([F = j/n]) = P([X = j]) = \binom{K}{j}\binom{N-K}{n-j} / \binom{N}{n}$ ; comme on a $E(X) = np$ et $\sigma(X) = \sqrt{np(1-p)\left(\frac{N-n}{N-1}\right)}$, on en déduit $E(F) = p$ et $\sigma(F) = \sqrt{p(1-p)\left(\frac{N-n}{N-1}\right)/n}$.

**Les approximations**
*(cf. tableaux des lois de probabilités usuelles et de leurs approximations)*
Le facteur $\frac{N-n}{N-1}$ est appelé en sondage "facteur de correction pour population finie" ; il est proche de 1 – *n*/*N* et donc proche de 1 lorsque le taux de sondage *n*/*N* (taille de l'échantillon sur taille de la population) est petit (< 0.1).

On montre que la loi hypergéométrique de paramètres *N*, *n*, p peut être approchée par la loi binomiale de paramètres *n*, *p* lorsque le taux de sondage *n*/*N* est petit (< 0.1) (on suppose que *N* et *K* tendent vers l'infini et que le rapport *K*/*N* tend vers un nombre *p*).

On montre que lorsque *p* est petit (< 0.1) alors la loi binomiale peut être approchée par une loi de Poisson de paramètre *n p* (on suppose que *p* tend vers 0, *n* vers l'infini et *n p* vers un réel $\lambda$).

Enfin, la loi hypergéométrique et la loi binomiale (resp. la loi de Poisson de moyenne $\lambda$) peuvent être approchées par une loi normale dès que la taille de l'échantillon *n* (resp. la moyenne $\lambda$) est suffisamment grande, $n \ge 30$ (resp. $\lambda > 10$).

**Théorème central limite**

Ces approximations reposent sur des théorèmes de convergence en loi, la convergence en loi de la loi binomiale vers la loi normale est un cas particulier du théorème central limite :

soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de v.a.r. i.i.d. admettant espérance et variance ; alors la loi de la moyenne d'échantillonnage centrée réduite converge vers la loi normale centrée réduite.

Formellement, si on pose $\mu = E(X_i), \sigma^2 = \text{var}(X_i)$ et $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ alors on a :

$E(\bar{X}_n) = \mu,\ \text{var}(\bar{X}_n) = \frac{\sigma^2}{n}$. La moyenne d'échantillonnage centrée réduite est la v.a.r. :

$\bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ et le théorème central limite s'écrit : $(\bar{X}_n^*) \xrightarrow{Loi} N(0;1)$

Cas particulier : schéma de Bernoulli
Si la loi de $X_i$ est la loi de Bernoulli de paramètre *p* (probabilité de succès) alors $\mu = p, \sigma = \sqrt{p(1-p)}$ et la moyenne d'échantillonnage n'est autre que la fréquence d'échantillonnage (fréquence de succès sur *n* expériences) notée $F_n$. La fréquence d'échantillonnage centrée réduite est alors la v.a.r. : $F_n^* = \frac{F_n - p}{\sqrt{p(1-p)/n}}$ dont la loi asymptotique est une loi normale centrée réduite (théorème central limite).

**Application numérique**

Dans les trois cas a) b) c), le taux de sondage *n* / *N* est négligeable ; la loi de la v.a.r. *X* peut être considérée comme une loi binomiale de paramètre *n* et *p* (où *p* est la proportion *K* / *N* de boules rouges dans l'urne) et la loi de probabilité de la fréquence de boules rouges dans l'échantillon *F* = *X* / *n* se déduit de celle de *X*.

Dans le cas b) la taille de l'échantillon *n* = 200 permet d'approcher la loi binomiale par la loi normale de même espérance et de même écart-type.

Dans le cas c) la proportion *p* étant faible, égale à 1 / 20, la loi binomiale peut être approchée par la loi de Poisson de paramètre *n p* soit ici 1.

a) $X \propto \mathcal{B}(20;0.3)$ ; $P(0.2 \le F \le 0.4) = P(4 \le X \le 8) \simeq 0.78$.

b) $F \propto \mathcal{N}(0.3;0.0324)$ ; $P(0.2 \le F \le 0.4) = 2P(0 \le U \le 3.086) \simeq 1$ (avec $U \propto \mathcal{N}(0;1)$ ).

c) $X \propto \mathcal{P}(1)$ ; $P(X > 1) \simeq 0.26$.

## 4. Statistique inférentielle, échantillonnage et estimation

La statistique inférentielle consiste à obtenir des informations sur les paramètres de la population (ici, proportion de boules rouges dans l'urne) à partir des résumés numériques observés sur un échantillon de taille *n* (ici, fréquence de boules rouges dans l'échantillon).
La partie *échantillonnage* consiste à déterminer les lois de probabilités des variables dites d'échantillonnage qui vont être utilisées dans la deuxième phase d'estimation des paramètres. Ici, il s'agit de déterminer la distribution de probabilité de la variable aléatoire $F_n$, fréquence d'échantillonnage de la catégorie d'intérêt. On en déduit *l'intervalle de fluctuation de* $F_n$ autour

de p à 95% de probabilité (nouveau programme de 2nde). Cette partie est déductive (calcul de probabilités).
La partie *estimation* consiste à partir d'UN échantillon observé à fournir une estimation par intervalle de confiance des paramètres de la population. Ici, c'est à partir d'une fréquence $f_n$ observée sur un échantillon de taille *n* que l'on estime *p* par intervalle de confiance à 95%. Cette partie est inductive (on infère un résultat d'un échantillon à la population avec une certaine probabilité d'erreur fixée à l'avance, ici 5%).

**Intervalle de fluctuation de la fréquence d'échantillonnage $F_n$ autour de *p* au niveau de probabilité 95%**

On lit sur la table de la loi normale centrée réduite, que si *U* est une v.a.r. $N(0\,;1)$ alors $P(-1.96 \le U \le +1.96) = 0.95$. On en déduit que si *X* est une v.a.r. $N(\mu_X\,;\sigma_X)$ alors $P(\mu_X - 1.96\,\sigma_X \le X \le \mu_X + 1.96\,\sigma_X) = 0.95$, donc, pour *n* suffisamment grand :

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \text{ et}$$

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \le F_n \le p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95.$$

Pour tout p de [0 , 1] on a $p(1-p) \le \frac{1}{4}$ et égalité pour $p = \frac{1}{2}$, on en déduit :

$$P\left(p - \frac{1}{\sqrt{n}} \le F_n \le p + \frac{1}{\sqrt{n}}\right) \ge 0.95 \text{ c'est-à-dire :}$$

on a l'inégalité $|F_n - p| \le \frac{1}{\sqrt{n}}$ avec une probabilité supérieure à 0.95.

**Estimation de *p* par intervalle à 95% de confiance centré sur $f_n$**

Lorsque l'on répète une épreuve de Bernoulli de paramètre *p* dans les mêmes conditions un grand nombre de fois *n*, on peut donc estimer la probabilité *p* de succès lors d'une expérience par la fréquence de succès $f_n$ sur les *n* épreuves (lettre minuscule car il s'agit d'une observation de la v.a.r. $F_n$ sur LA série de *n* épreuves réalisées). Il s'agit du *deuxième lien entre fréquence et probabilité* introduit dans les programmes sous le nom *d'approche fréquentiste de la probabilité.*

De la proposition "on a l'inégalité $|F_n - p| \le \frac{1}{\sqrt{n}}$ avec une probabilité supérieure à 0.95", on déduit la proposition "on a l'inégalité $|f_n - p| \le \frac{1}{\sqrt{n}}$ à 95 % de confiance". Quand on ne peut plus parler de *probabilité* (puisqu'il n'y a pas de v.a.r. mais une observation d'une v.a.r.) on parle de *confiance.*

On peut donc estimer *p* par l'intervalle à 95 % de confiance : $\left[f_n - \frac{1}{\sqrt{n}}\,;\, f_n + \frac{1}{\sqrt{n}}\right]$.

Dans le cadre du tirage d'un échantillon à probabilités égales avec remise de taille *n* dans une population finie (approche sondage), à chaque échantillon on associe la fréquence de

succès $f_n$ et l'intervalle $\left[ f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}} \right]$, on peut dire qu'au moins 95 % des échantillons fournissent un intervalle $\left[ f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}} \right]$ qui contient la proportion *p*.

Dans le cadre expérimental, on pourra seulement conclure que "la probabilité *p* de succès lors d'une expérience appartient à l'intervalle $\left[ f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}} \right]$ à 95 % de confiance, donc avec un risque de se tromper de 5 %".